Streaming Algorithms for Set Cover

Piotr Indyk
With*: Sepideh Mahabadi, Ali Vakilian*

Set Cover

- Input: a collection *S* of sets $S_1 \ldots S_m$ that covers $U=\{1\ldots n\}$
  - I.e., $S_1 \cup S_2 \cup \ldots. \cup S_m = U$
- Output: a subset *I* of *S* such that:
  - *I* covers U
  - |*I*| is minimized
- Classic optimization problem:
  - NP-hard
  - Greedy ln(n)-approximation algorithm
  - Can't do better unless P=NP (or something like that)

Streaming Set Cover [SG09]

- Model
  - Sequential access to S_1, S_2, …., S_m
  - One (or few) passes, sublinear (i.e., o(mn)) storage
  - (Hopefully) decent approximation factor
- Why ?
  - A classic optimization problem (see previous slide)
  - Several ``big data'' uses
  - One of few NP-hard problems studied in streaming
    - Other examples: max-cut, sub-modular opt, FPT

The ``Big Table''

Result	Approximation	Passes	Space	R/D
Greedy	ln(n)	1	O(mn)	D
Greedy	ln(n)	n	O(n)	D
[SG09]	O(logn)	O(logn)	O(n logn)	D
[ER14]	$O(n^{1/2})$	1	$\tilde{O}(n)$	D
[DIMV14]	$O(4^{1/\delta}\rho)$	$O(4^{1/\delta})$	$\tilde{O}(mn^{\delta})$	R
[CW]	n^{δ}/δ	$1/\delta-1$	$\tilde{\Theta}(n)$	D
[Nis02]	log(n)/2	O(logn)	$\Omega(m)$	R
[DIMV14]	O(1)	O(logn)	$\Omega(mn)$	D

[IMV]	$O(\rho/\delta)$	$O(1/\delta)$	$\tilde{O}(mn^{\delta})$	R
[IMV]	1	$1/2\delta-1$	$\tilde{\Omega}(mn^{\delta})$	R
[IMV]	1	$1/2\delta-1$	$\tilde{\Omega}(ms)$	R
[IMV]	3/2	1	$\Omega(mn)$	R

A few observations: algorithms

Greedy	ln(n)	1	O(mn)	D
Greedy	ln(n)	n	O(n)	D
[SG09]	O(logn)	O(logn)	O(n logn)	D
[ER14]	O(n)	1	$\tilde{O}(n)$	D
[DIMV14]	$O(4^{1/\delta}\ \rho)$	$O(4^{1/\delta})$	$\tilde{O}(mn^{\delta})$	R
[CW]	n^{δ}/δ	$1/\delta-1$	$\tilde{\Theta}(n)$	D
[IMV]	$O(\rho/\delta)$	$O(1/\delta)$	$\tilde{O}(mn^{\delta})$	R

- Most of the algorithms are deterministic
- All of the algorithms are ``clean''

A few observations: lower bounds

[Nis02]	log(n)/2	O(logn)	$\Omega(m)$	R
[DIMV14]	O(1)	O(logn)	$\Omega(mn)$	D
[CW]	n^{δ}/δ	$1/\delta-1$	$\tilde{\Theta}(n)$	D
[IMV]	1	$1/2\delta-1$	$\tilde{\Omega}(mn^{\delta})$	R
[IMV]	3/2	1	$\Omega(mn)$	R

Algorithm

[IMV]	$O(\rho/\delta)$	$O(1/\delta)$	$\tilde{O}(mn^{\delta})$	R

- Approach: "dimensionality reduction"
  - Covers all but $1/n^{\delta}$ fraction of elements using ρ*k sets (k=min cover size)
  - Uses $\tilde{O}(mn^{\delta})$ space
  - Two passes
- Repeat $O(1/\delta)$ times:
  - $O(1/\delta)$ passes
  - $O(\rho/\delta)$ approximation

Dimensionality reduction:

- Covers all but $1/n^{\delta}$ fraction of elements
- Uses mn^{δ} space
- Two passes

- Suppose we know k=min cover size
- Pass 1:
  - For each set S_i , select S_i if it covers $\Omega(n/k)$ elements
  - Compute V=set of elements not covered by selected sets
  - Fact: each not-selected set covers $O(n/k)$ elements in V
- Select a set R of $kn^{\delta} \log m$ random elements from V
- Pass 2:
  - Store all sets projected on R
  - Compute a ρ-approximate set cover I'
  - Fact [DIMV14, KMVV13]: I' covers all but $1/n^{\delta}$ fraction of V
- Report sets found in Pass 1 and Pass 2

Dimensionality reduction: space accounting

- Suppose we know k=min cover size $* \log n$
- Pass 1:
  - For each set S_i , select S_i if it covers $\Omega(n/k)$ elements n
  - Compute V=set of elements not covered by selected sets
  - Fact: each not-selected set covers O(n/k) elements in V
- Select a set R of $kn^{\delta} \log m$ random elements from V
- Pass 2:
  - Store all sets projected on R
  - Compute a ρ-approximate set cover I'
  - Fact [DIMV14, KMVV13]: I' covers all but $1/n^{\delta}$ fraction of V
- Report sets found in Pass 1 and Pass 2

$m*(n/k)*|R|/n$
$=m*n^{\delta} \log m$

Lower bound: single pass

[IMV]	3/2	1	Ω(mn)	R

- Have seen that O(1) passes can reduce space requirements
- What can(not) be done in one pass ?
- We show that distinguishing between k=2 and k=3 requires Ω(mn) space

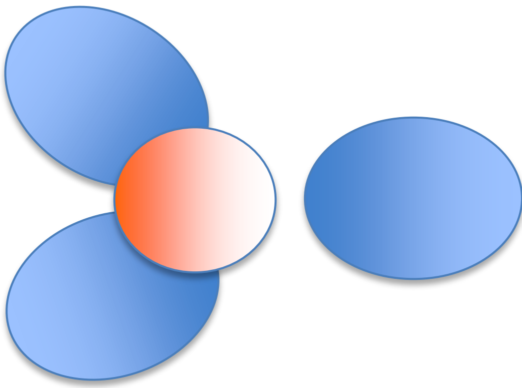
Proof Idea

- Two sets cover U iff their complements are disjoint
- Consider two following one-way communication complexity problem:
  - Alice: sets $S_1 \ldots S_m$
  - Bob: set S
  - Question: is S disjoint from one of S_i's ?
- Lemma: the randomized one way c.c. of this problem is $\Omega(mn)$ if error prob. is 1/poly(m)

Proof idea ctd.

- Lemma: the one way c.c. of this problem is $\Omega(mn)$ if error prob. is 1/poly(m).
- Proof:
  - Suppose S_i's are selected uniformly at random
  - We show that there exist poly(m) sets S such if Bob learns answers to all of them, he can recover all S_i's with high probability

Proof idea ctd.



- Bob's queries:
  – poly(m) random "seed" queries of size c log m for some constant c>0
  – For each sees query S, all "extension" queries of the form $S \cup \{i\}$
- Recovery procedure
  – Suppose that a seed S is disjoint from *exactly* one S_i (we do not know which one)
    - Call it a ``good seed'' for S_i
  – Then extension queries recover the complement of S_i
- poly(m) queries suffice to generate a good seed for each S_i

Lower bound: multipass

[IMV]	1	$1/2\delta-1$	$\tilde{\Omega}(mn^{\delta})$	R
[IMV]	1	$1/2\delta-1$	$\tilde{\Omega}(ms)$	R

- Reduction from Intersection Set Chasing [Guruswami-Onak'13]
- Very "brittle", hence works only for the exact problem

Conclusions

Result	Approximation	Passes	Space	R/D
Greedy	$\ln(n)$	1	$O(mn)$	D
Greedy	$\ln(n)$	n	$O(n)$	D
[SG09]	$O(\log n)$	$O(\log n)$	$O(n \log n)$	D
[ER14]	$O(n^{1/2})$	1	$\tilde{O}(n)$	D
[DIMV14]	$O(4^{1/\delta} \rho)$	$O(4^{1/\delta})$	$\tilde{O}(mn^{\delta})$	R
[CW]	n^{δ}/δ	$1/\delta-1$	$\tilde{\Theta}(n)$	D
[Nis02]	$\log(n)/2$	$O(\log n)$	$\Omega(m)$	R
[DIMV14]	$O(1)$	$O(\log n)$	$\Omega(mn)$	D

[IMV]	$O(\rho/\delta)$	$O(1/\delta)$	$\tilde{O}(mn^{\delta})$	R
[IMV]	1	$1/2\delta-1$	$\tilde{\Omega}(mn^{\delta})$	R
[IMV]	1	$1/2\delta-1$	$\tilde{\Omega}(ms)$	R
[IMV]	3/2	1	$\Omega(mn)$	R